# 高頻度データを用いたニュースと韓国株式の分析:ニュース分類の条件探索

## Analysis of News and Stock Returns in Korean Markets Using High Frequency Data: Exploring Conditions of Classification Algorithm

尹聖在[*1] Sungjae Yoon　菅愛子[*1] Aiko Suge　高橋大志[*1] Hiroshi Takahashi

[*1] 慶應義塾大学大学院経営管理研究科
Graduate School of Business Administration, Keio University

News articles play an important role in financial markets. This study analyzes the relationship between news articles and stock price fluctuations using high frequency trading data in Korean stock markets. Especially, we analyze differences in market reactions according to languages of news articles. In order to understand the influences of news articles, this study explores conditions of Long Short Term Memory (LSTM) models that classify news articles.

## 1. はじめに

フィンテックなど金融分野における技術の進展を背景とし，資産運用への関心が高まっている．資産運用分野において，新たな情報技術を用いた分析などが盛んに報告される中，金融市場において発信されるニュースは，資産価格に関する重要な情報の一つに挙げられる．効率的な市場において，情報は資産価格に速やかに反映されるという理論背景において，日々発信されるニュースと株式市場の関連性について，これまで数多くの実証研究がなされてきた．近年では，ニュース記事の内容を定量化・指標化する試みがなされ，それらと投資家行動に及ぼす要素とを関連付ける研究もされている．

我々は以前，様々な言語のニュースが株価に与える影響を韓国株式市場において分析した（図 1）[Yoon, 2017] [Yoon, 2018]．更にその中で，Long Short Term Memory (LSTM)ブロック [Hochreiter & Schmidhuber, 1995] [Hochreiter & Schmidhuber, 1997] を用いた深層学習モデルにて，英語ニュースによる韓国株式のリターン分類を報告した．そのモデル作成時のハイパーパラメータ探索の試みを本稿で報告する．

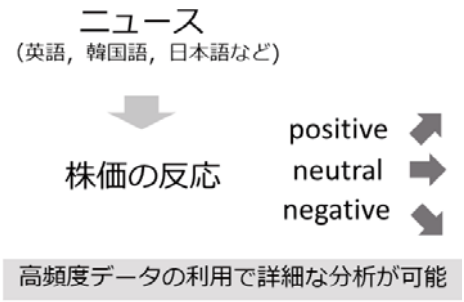


図 1:ニュースガ株価に与える影響

## 2. 目的

機械学習モデルの学習において，良いモデルを作るためにハイパーパラメータ（外部から与えるモデルのパラメータ）のチューニングが必要となる．良いモデルの一例として，精度の高いモデルが挙げられる．目的の精度をもとめてモデル学習を繰返す際，選択するアルゴリズム，データ量，計算環境などによって，一回の試行に時間を要することも多い．そこで効率よく探索を行うため，様々な試みがなされており，ランダムサーチ，グリッドサーチ，ベイズ最適化，などが知られている．

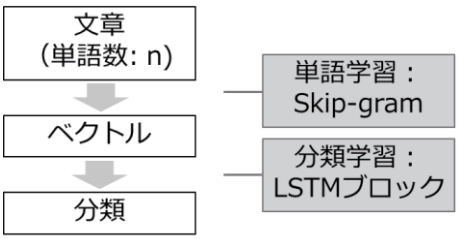


図 2:モデルの概要

深層学習では，多数のパラメータが存在する．例えば筆者ら用いたモデルは図 2 のように要約されるが，そこで用いるハイパーパラメータとして例えば，学習係数，隠れ層のユニット数，層の段数，epoch 数，dropout，batch size などが挙げられる．

本稿ではパラメータ探索の始めとして，どの段階におけるパラメータの影響が大きいのか，ベクトル（入力段階），分類学習の段階，で大きく分けパラメータの変更を行い，分類結果（精度）の変化を観察した．

## 3. データ

### 3.1 ニュースデータ

ニュースデータは韓国株式市場に上場している時価総額上位 5 社（2012 年末時点）について Thomson Reuters から抽出した．2013 年ニュースをモデル作成用（学習用）に用い，2014 年ニュースを予測に用いた．ニュース発信言語は英語の他，韓国語，日本語，中国語など多岐に渡るが，本稿では英語ニュースを扱う．

### 3.2 株価データ

Thomson Reuters の高頻度データから，上記で抽出したニュース発信前後の株価を 15 分ごとに取得した．ニュース発信 30 分前に比較し，ニュース発信 30 分後に株式が 0.3%以上値上がりしたものをポジティブ，0.3%以上値下がりしたものをネガティブ，それら以外をニュートラルなニュースであると分類した．結果，表 1 のようなデータセットが得られた．

連絡先:菅愛子，慶應義塾大学大学院経営管理研究科，aikosuge@keio.jp

表 1:ニュースデータの分布

| | ネガティブ | ニュートラル | ポジティブ | 計 |
|---|---|---|---|---|
| 2013 年 | 15.9% | 63.5% | 20.6% | 908 |
| 2014 年 | 19.3% | 62.8% | 17.9% | 643 |

## 4. 方法

基本モデルを設定し, その中のハイパーパラメータを変化させ, 精度の変化を検証した.

### 4.1 基本モデルの設定

図 2 に示すようにニュース分類をセットした. すなわち, 文章(単語)のベクトル化には skip-gram 法 [Mikolov, 13] を, 株価文章分類の学習には LSTM ブロックを用いた. 2013 年ニュースデータ(908 サンプル)を訓練データとテストデータに分け(9:1), モデル学習には訓練データを用いた.

### 4.2 パラメータの変更

入力段階においては, 文章をベクトル化する際に用いている skip-gram のパラメータ, すなわち学習に用いる単語数やベクトルの次元, 学習効率(initial learning rate: initial LR, epoch 数)を変化させた. また, ニュース分類段階では LSTM ブロックの学習効率(initial LR, epoch 数)を変化させた.

### 4.3 結果の検証

訓練データとテストデータの, 精度の比較を行った. また, いくつかのモデルにおいて, 2014 年ニュースデータを用い, 株価リターン分析を行い, 分類予測を行った. なお, 結果は 20 回の学習を行い, その精度を平均して示している.

## 5. 結果

表 2 にパラメータ探索の結果を一部示す. なお, 結果は 20 回の学習を行い, その精度を平均して示している.

表 2:条件による精度の変動

| | 入力側 | LSTM ブロック | | 検証 | |
|---|---|---|---|---|---|
| 条件 | 単語数: n | Initial LR | Epoch 数 | 訓練精度 | 検証精度 |
| 1 | 100 | 0.01 | 100 | 63.6% | 63.3% |
| 2 | 100 | 0.01 | 200 | 63.4% | 64.4% |
| 3 | 100 | 0.01 | 300 | 63.4% | 64.4% |
| 4 | 100 | 0.1 | 200 | 70.7% | 63.3% |
| 5 | 200 | 0.01 | 200 | 63.3% | 63.6% |

今回の結果では, 入力する単語数 n が精度に与える影響は小さいことが分かった. しかしながら, 今回用いたテキストデータの最長は 900 単語を超えるものもあり, 今後, 入力条件について詳細に分析していきたい. また, 分類学習における initial LR を 0.1 以上にすると, 訓練データが示すモデル精度と検証データが示す精度で乖離が見られ, 過学習が起きていることが分かった. 精度が 60%程度にとどまっている原因として, サンプル数の少なさや, データの不均衡分布が考えられる. 今後はその点を調整して学習モデルの改善に取り組む予定である.

最後に表 2 の条件 2 について, 2014 年のニュースデータを用い分類を行い, 実際の株価リターン(平均)を算出した. その分類結果を表 3 に示す. この時の予測精度は 55.7%となった.

表 3:学習モデルによる 2014 年ニュース分類(条件 2)

| 予測 | ネガティブ | ニュートラル | ポジティブ | 計 |
|---|---|---|---|---|
| 分布 | 17.0% | 73.7% | 9.3% | 643 |

今回はデータが限られており, 別の期間などのデータを用いて, モデルや精度の検証をする必要がある.

## 6. まとめ・課題

ニュースによる株価分類を深層学習にてモデル化した際のハイパーパラメータについて探索した. 数多くのパラメータが存在するが, 今回は入力単語数と, モデル学習における initial RL と epoch 数について変動させ, モデル精度に与える影響を観察した. 過学習が起きない一定の条件を見出したものの, 全体の精度は 60%程度にとどまった. 今後, 更なる精度向上のためにモデルやパラメータ設定が課題である.

また深層学習以外の分類法との比較も望まれる.

## 参考文献

[Hochreiter & Schmidhuber, 95] Sepp Hochreiter and J¨urgen Schmidhuber: Long Short Term Memory, Technical Report FKI-207-95, Technische Universit¨at M¨unchen, M¨unchen, August 1995.

[Hochreiter & Schmidhuber, 97] Sepp Hochreiter and J¨urgen Schmidhuber: Long Short Term Memory, Neural Computation, 9(8):1735–1780, November 1997.

[Mikolov, 13] Mikolov, T., Chen, K., Corrado, G., and Dean, J.: Efficient estimation of word representations in vector space, ICLR, 2013.

[Yoon, 17] Yoon, S.J., Suge, A., and Takahashi, H.: Analyzing the Influence of News Articles on the Korean Stock Market with High Frequency Trading Data, JSAI International Symposia on AI, Workshop 3: Artificial Intelligence of and for Business (AI-Biz 2017), 2017.

[Yoon, 18] Yoon, S.J., Suge, A., and Takahashi, H.: The Analysis on Korean Stock Market by High Frequency Data and Text Classification Using LSTM, JSAI Special Interest Group on Business Informatics (SIG-BI #8), 2018.